Résumé

Introduction

comment produire des informations normalisées et exploitables à partir d'un corpus textuel semi-structuré ?

Première partie

Du document numérisé au XML-TEI : nature du corpus, structure des documents et méthode de production des données

Chapitre 1

Présentation du corpus

Ce chapitre est dédié à une présentation des documents traités dans le cadre du projet MSS : nature, quantité de documents (et d'entrées individuelles), dates et différents types de catalogues vendus. On pourra également représenter la répartition des ventes par an grâce aux graphiques produits pour le site web avec Plotly. Cette partie s'appuie sur les mémoires effectués par d'ancien.ne.s stagiaires de Katabase, qui ont déjà beaucoup analysé la nature et les enjeux du corpus (Lucile Rondeau du Noyer, par exemple).

Chapitre 2

Production des données : OCRisation et structure des documents traités

Cette partie s'attache autant à présenter le processus d'OCRisation (qui est déjà bien établi et ne constitue pas le cœur de mon stage) que la structure des documents. Alors que le chapitre d'au dessus s'intéresse au catalogues dans leur ensemble, ici, on étudie le corpus au niveau de la page et de l'entrée individuelle. En effet, l'OCRisation repose sur la segmentation, et donc sur l'établissement d'une structure « abstraite » d'une page (c'est-à-dire, d'un découpage de la page en zones).

Chapitre 3

Du texte à la TEI : méthode de transformation des documents OCRisés en fichiers XML-TEI valides

Après une étape d'OCRisation via eScriptorium, le texte extrait des PDFs peut être exporté soit en texte brut, soit en XML Page ou Alto. Ces formats s'attachent à garder une relation entre le XML et le document numérisé (les zones de texte sont indiquées, chaque ligne est dans une balise...). Cependant, l'unité intellectuelle centrale à la suite du projet, ce n'est pas la page numérisée, mais l'entrée de catalogue. Un format plus complexe que le XML d'eScriptorium est donc nécessaire. Assez logiquement, la suite du projet s'appuie sur une traduction des catalogues en TEI. Jusqu'à maintenant, cette transformation était faite par GROBID-Dictionnaries et des feuilles XSL. Le projet GROBID-Dictionnaries étant maintenu de façon assez opaque, le choix a été fait de remplacer cet outil par une solution plus simple (qui ne passe pas par le *machine learning*) et modulable. En s'appuyant sur la nature « semi-structurée » du corpus, il est possible de séparer les différentes parties du texte à en identifiant des séparateurs. Ce chapitre s'intéresse donc à la transformation de Alto vers la TEIà l'aide d'*expressions régulières*. En changeant d'unité intellectuelle – ou en choisissant de privilégier une unité au sein des catalogues plutôt qu'une autre (page, série d'entrées...) –, on prend de la distance et on interprète le catalogue papier.

Deuxième partie

Normalisation, enrichissements et extraction d'informations : une chaîne de traitement pour des données semi-structurées

Chapitre 4

Homogénéiser et normaliser un corpus complexe

Ce chapitre s'intéresse à la manière dont les fichiers `TEI` sont traités afin de pouvoir en extraire des informations. C'est directement grâce la structure des entrées (et grâce à la nature « semi-structurée » des catalogues) qu'est possible le traitement automatisé des documents. Cette partie s'attache également à rappeler les questions de recherche qui sous-tendent la normalisation des documents (ajouter plus de structure au document `TEI` pour l'exploiter plus facilement, uniformiser la notation des tailles et des dimensions des documents...). Cette partie sera probablement relativement brève, puisque l'étape de normalisation des données a déjà été faite par d'autres stagiaires et documentée d'autres mémoires.

Chapitre 5

Extraction d'informations

double enjeu (garantir info historique et utilisation contemporaine)

Chapitre 6

Vers une étude des facteurs déterminant le prix des documents : alignement des entrées du catalogue avec Wikidata et exploitation de données normalisées

Ce chapitre est construit autour d'une question de recherche : comment produire des informations exploitables pour une étude économétrique à partir d'un corpus textuel semi-structuré ? Un des objectifs du projet est de faire l'étude des facteurs déterminant le prix d'un manuscrit. Pour faire cette étude, il faut obtenir, pour chaque entrée du catalogue, un certain nombre d'informations normalisées. Le travail d'extraction de données présentes dans les catalogues a déjà été fait par de précédent.e.s stagiaires. Ces données sont principalement quantitatives : prix des manuscrits, dimensions et nombre de pages, date de création. Il est nécessaire de compléter les informations par des données qualitatives et d'enrichir les manuscrits depuis des sources extérieures. Pour ce faire, il a été choisi d'aligner le nom des auteur.ice.s des manuscrits avec des identifiants Wikidata ; dès lors que l'on a un identifiant Wikidata, il est possible de récupérer automatiquement des informations sur les personnes. Le choix de travailler uniquement sur les noms, et non sur la description des documents, a deux motivations :

— Les noms de personnes (et la manière dont elles sont décrites) constituent la partie la plus normalisée des documents. La description des manuscrits est plutôt en "texte libre". Dans la continuité avec le reste du projet, nous sommes resté dans une approche « basse technologie » (*low-tech*), qui consiste à s'appuyer majoritairement sur des solutions techniquement simples. C'est pourquoi nous avons préférer traiter les noms avec des tables de correspondance et des *expressions*

régulières, plutôt que de faire du TAL sur la description des documents.

— Toutes les informations « simples » (informations quantitatives facilement normalisables : dates etc.) ont déjà été extraites des descriptions des manuscrits.

Ce travail d'enrichissement a été fait en deux temps.

La première étape, et la plus difficile, est l'alignement avec Wikidata. Cela demande d'extraire un ensemble d'informations à partir du nom de la personne et de la description de celle-ci (nom, prénom, titre de noblesse, occupation, dates de vie et de mort...). À partir de ces informations, stockées dans un dictionnaire, un algorithme construit successivement différentes chaînes de caractères à rechercher sur l'API de Wikidata. L'objectif est que le premier résultat recherché sur Wikidata soit correct. Sur un jeu de test, le score F1 obtenu est de 68%. Un relecture « manuelle » des résultats est donc nécessaire.

La deuxième étape, nettement plus simple, consiste à lancer des requêtes Wikidata sur les identifiants récupérés afin d'enrichir les entrées d'informations sur les auteur.ice.s des manuscrits (cette partie du travail est encore en cours).

Troisième partie

Après la TEI : l'application web *Katabase*, interface de diffusion des données

Chapitre 7

Utiliser une interface web et présenter le corpus différemment

Ici, on s'intéresse à la manière dont le site web de Katabase permet de proposer une visualisation différente du corpus :

- création de différentes manières de naviguer dans le corpus et mise en lien entre le corpus et le projet Katabase
- création d'une interface de recherche et de réconciliation des manuscrits, qui recoupe les différentes ventes afin d'identifier si un manuscrit a été vendu plusieurs fois
- visualisations de données et production de graphiques interactifs sur la page d'index des catalogues et pour chaque catalogues

Les deux premières parties ont été réalisées par Alexandre Bartz et décrites dans son mémoire. Elles ne concernent pas le cœur du projet. J'y reviens cependant pour mettre en avant que la création d'une interface web implique l'éloignement du document originel. Cet éloignement est théorique : le « catalogue » n'est plus l'unité intellectuelle dominante, alors qu'il restait l'un des critères structurants des fichiers `TEI` (un catalogue représentant un fichier). Sur le site web, on peut accéder directement aux éléments vendus, sans avoir à passer par les catalogues. L'éloignement du document originel est enfin technique, puisqu'on abandonne totalement la `TEI` au profit de formats de présentation (`HTML`), tout en s'appuyant largement sur des formats « techniques » et pratiques (le `JSON`, qui n'est pas un format de conservation, mais simplement un moyen de rendre des données produites rapidement accessibles).

Dans cette partie, je m'intéresserai plus spécifiquement aux problématiques techniques et scientifiques liées à la visualisation de données, qui ont constitué une des missions centrales de mon stage :

- Qu'est-ce que l'on cherche à montrer avec ces visualisations ?
- Comment intégrer ces visualisations de manière "élégante" au site ? Par exemple, 7 types de graphiques sont produits pour l'index des catalogues. Il faut donc

trouver un moyen de tous les présenter, sans pour autant alourdir excessivement la page. La visualisation doit également être pensée en concertation avec le reste du design du site web (largement mis à jour pendant mon stage), et s'intégrer à sa charte graphique.

— Comment construire ces visualisations ? Comment faciliter la lecture d'informations complexes ? Quel impact les visualisations ont-elles sur la perception des informations ? Pour ces problématiques techniques, je pense (entres autres) m'appuyer sur le mémoire de Ségolène Albouy.

Chapitre 8

Visualisation et interfaces : un problème pour les humanités numériques ?

Ce chapitre de conclusion, plus théorique, revient sur les débats actuels concernant le lien entre interfaces graphiques, visualisation de données et humanités numériques. En s'appuyant (entre autres) sur les théories de Johanna Drucker et Anthony Masure, il s'attache à réintégrer les questions « visuelles » propres aux humanités numériques à un contexte plus large. Le chapitre revient sur les origines des interfaces graphiques (et donc sur les objectifs implicites qui structurent nos interactions avec les machines) ; il cherche à remettre la visualisation dans les humanités numériques en lien avec une tendance globale à la visualisation – tendance qui vient du monde de l'entreprise. Du fait de la quantité de significations implicites sous-jacentes à la construction d'interfaces, il est, je pense, nécessaire d'avoir une approche théorique et critique du rapport des humanités numériques aux problématiques de design et aux « dispositifs » (Agamben) que sont nos outils de travail.

Table des figures

Liste des tableaux

Table des matières